Improving the Portability of Machine Learning Models for NLP by Mixing Linguistics from Multiple Social Media Platforms

Daniel Furman and **Benjamin Fell** and **Melissa Licari**
University of California, Berkeley,
School of Information
{daniel-furman; benjamin.fell; melissa-licari}@berkeley.edu

Abstract

Domain shift often hinders the predictive performance of machine learning models where it counts most, on unseen data. However, social media datasets in NLP are in general inflexible to domain shift as they are commonly sourced solely from Twitter, which fails to capture the variation of natural language that exists across different social media platforms. Here, we examined the potential of multi-platform mixing for domain adaptation by combining Instagram captions in equal proportion to Tweets for two authorship analysis tasks. The resulting SVM and LR classifiers saw significant boosts in performance when compared to otherwise identical models constructed entirely from Tweets (6% average F1 increase), as measured in cross-domain testing on Facebook captions.

1 Introduction

In statistical learning, the datasets employed for training and testing often originate from distinct distributions, a phenomenon referred to as domain shift. Domain shift is a crucial consideration because it can occur in any given supervised setting and hinders the portability of statistical learners, as real-world testing data differs starkly from typical training sets (Plank, 2016; Peters et al., 2017). While innovations in domain adaptation have emerged, including methods in data augmentation and neural self-learning (Zhang et al., 2018; Ramponi and Plank, 2020; Butt et al., 2021; Baevski et al., 2022), domain shift remains a central concern in machine learning for natural language processing (NLP).

Domain shift is common in NLP due to the emergence of massive, labelled corpora sourced from canonical standards (Plank, 2016). The variation of natural language type between training data, often of canonical type, and deployment data typically involves dimensions such as socio-demographics, language, character and word length, sentence type, and genre (Plank, 2016). Examples of such standards include the seminal Penn Treebank, a three-year collection of Wall Street Journal articles introduced in the 1980's, and the more recently developed GLUE and SuperGLUE benchmarking datasets. And, for social media corpora, Twitter has now emerged as the canonical standard.

A social media platform is defined here as a medium where content is shared among the general public, such as Twitter, Instagram, Facebook, Reddit, and Pinterest. In this study, we argue that mixing Tweets with captions from less frequently mined, hitherto under-used social media platforms improves the quality of data captured in the training set and ultimately leads to the development of more robust statistical learning models. The potential of the multi-platform mixing technique was examined herein with a custom dataset of public, verified celebrity captions drawn from Twitter, Instagram, and Facebook for two cross-domain tasks in text classification for authorship analyses. Our results suggest that multi-platform mixing indeed improves the portability of classifiers for these tasks, setting concrete expectations for domain-mixing techniques in machine learning for NLP.

2 Related Work

2.1 Authorship Analysis

Authorship analysis refers to predicting the identity or type of author from natural language. The field has been an active area of research in NLP for decades (M. L. Jockers, 2010; Sari et al., 2018) and was chosen for our experiments due to the frequency at which such models are applied to social media corpora. Authorship analysis models are commonly trained on long pieces of text such as journals, books, and articles, leading to an problem when using social media data, in that captions are generally very short and less stylometric features can be collected per document. Authorship analysis on social media thus requires a large amount of training data to reach reasonable classification accuracy (Anderson Rocha, 2016).

Wiegmann advanced the state-of-the-art in authorship analysis for social media by exploring gender profiling within a large corpus of celebrity Tweets (Wiegmann et al., 2019). The authors ensured each account was verified, to ensure privacy protocols were met. In addition, the PAN conference also often push the state-of-the-art for authorship analysis tasks through hosting modeling competitions. For example, the PAN19 authorship attribution challenge was to predict the author of fanfiction text across several domains and focused on open-set attribution (Kestemont et al., 2019). The best performing solution used Singular Value Decomposition and an ensemble of Support Vector Machine (SVM) classifiers for modeling architectures. In the PAN18 celebrity profiling challenge, the best performing solution for prediction of gender, age, occupation and fame level from Tweets employed an ensemble of Logistic Regression (LR) and SVM classifiers (Radivchev et al., 2019). Radivchev also tested neural networks but found that these models were less effective in comparison.

2.2 Domain Adaptation in NLP

Reducing domain shift in machine learning is referred to as domain adaptation. Overall, these methods are directly aimed at enforcing the distributions of the training and testing data to be as similar as possible (Luo et al., 2019; Ramponi and Plank, 2020). Common methods in NLP include instance weighting, data augmentation, domain-adversarial neural networks, and transfer learning (Zhang et al., 2018; Ramponi and Plank, 2020; Butt et al., 2021; Baevski et al., 2022) (pers. comm., Bamman, 2021; A1). For example, MixUp is a data augmentation method that improves the portability of models by reducing memorization and sensitivity to domain shift. The technique trains a deep learning model on convex combinations of data instances, which regularizes the model to favor more linear behavior as opposed to non-linear, overfitting behavior. Multi-platform mixing is similarly aimed at improving model generalization by improving robustness to adversarial, out-of-domain examples. Unlike MixUp, which is confined to neural network training, multi-platform mixing with social media platforms or mixing multiple types of data more generally is a method applicable to most statistical learning models and tasks. By providing more data sources, platform-mixed models can more effectively avoid potential situations where a single platform offers limited data and those where the size of data available for training on a set of authors are substantially different.

Employing multi-platform mixing with Instagram and Twitter data has been previously shown to improve domain adaptation in personality classification for recommendation systems (Skowron et al., 2016). Skowron trained multi-modal models with Twitter and Instagram posts and found that the multi-platform mixed models performed significantly better when validated on three benchmarks, one of which included Facebook data as a form of cross-domain testing. Our study was motivated by Skowron's work and aims to extend their results to more common tasks in NLP, such as those without multi-modal elements. And, in employing custom datasets curated specifically for this research, we aim to further isolate the impact of multi-platform mixing on model generalization.

3 Methods

3.1 Data

A custom dataset of social media captions composed solely of natural language was constructed by scraping posts from 237 celebrity social media accounts, which were selected among the 300 most followed accounts on Twitter. Critically, each Twitter account in the corpus corresponded to a verified Instagram and Facebook profile.[1] We ensured that all profiles included had a verified, public status so that no privacy concerns were violated, as was performed by Wiegmann (Wiegmann et al., 2019).

A multi-platform mixed dataset of equal proportion Instagram and Twitter captions was crafted for comparison against a second dataset of purely Tweets, which were equal in overall size. Thirdly, a cross-domain testing set of Facebook captions was created for validation. These datasets were sampled for visual inspection in A2 (Figure 2). They were then ran through equivalent pre-processing and featurization pipelines, which were motivated by the techniques used by Radivchev in PAN18 (Radivchev et al., 2019) and detailed in A3. While equivalent in all other ways, the sole difference

[1] Web-scrapers employed include Twitter Developer Portal tools and RapidAPI tools.

between the train and test pipelines was our taking lower-cased captions in pre-processing the Facebook test datasets, so to induce further domain shift for our experiments on model generalization.

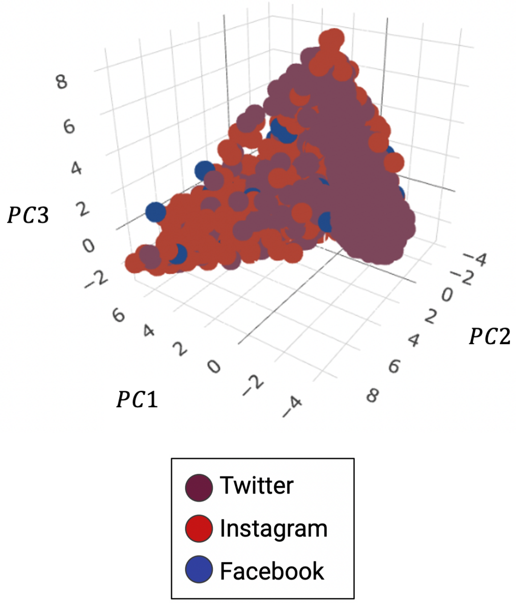


Figure 1: Social media captions were featurized by averaging BERT embeddings across word tokens from the Base (Cased) final layer and then visualized in 3-d Principal Component Space. Twitter captions clustered most tightly in comparison to Instagram or Facebook linguistics, possibly a result of the more rigid character limits therein among other differences in domain between platforms.

For authorship attribution, the Instagram and Twitter captions were taken as individual posts and piped into class balanced datasets for training (372 rows) while the Facebook data were piped into a class balanced testing set (60 rows). In contrast, for gender profiling, captions were concatenated with "sep" tokens per profile. We took a total of 80 percent of the accounts for training and reserved the remaining 20 percent for testing. The Instagram and Twitter captions were likewise piped into datasets for training (372 rows), while the Facebook posts was separated into a class balanced testing set (50 rows). However, the training sets for gender profiling were distinct from the authorship attribution datasets in that they contained a 37.6 percent class imbalance between female and male labels. As a result, we employed class balance weighting for this task. Only female and male labels were considered so to preserve as much class balance as possible. Yet, given the fluidity of gender, it is possible that the quality of the gender labels will decrease over time. Lastly, the entirety of the roughly 50,000 post-processed captions were visualized per individual post by transforming averaged BERT-Base (Cased) word token embeddings into 3-D Principal Component space, which provided evidence for the notion that different social media platforms indeed represent distinct domains with differing natural language type (Figure 1).

3.2 Modeling

We trained and evaluated two shallow learning models for our authorship analysis experiments on multi-platform mixing, Support Vector Machines (SVMs) and Logistic Regressors (LRs). These algorithms had similar performance during cross-validation fitting and were superior to other common machine learners. Models such as XGBoost, Random Forest, BERT-Base (Cased), and Google's Natural Language AutoML incurred anywhere from roughly a 4% to 15% drop in F1 score on cross-validation folds in comparison. The Term Frequency - Inverse Document Frequency (TF-IDF) featurization method was employed for training the shallow learning models by taking up to the 10,000 most frequent bi-grams in the given vocabulary, the same parameters employed by Radivchev (Radivchev et al., 2019). BERT-Base (Cased) embedding features were also added to the authorship attribution models, which boosted F1 performance by roughly 10% for each model (A4).

Training sets for the mixed-platform and single-platform models were then separated into 5 class-stratified folds for the purpose of cross-validation tuning. Models were fit by taking one fold per iteration as the validation set and the others as the training set. Each was tested with a variety of hyper-parameters by using the grid search technique. And, as previously mentioned, class balance weighting was employed for gender profiling to overcome the class-imbalance in the training set. The fold with the best F1 performance was the selected for both the mixed-platform dataset and the Tweet-only dataset for the purposes of re-fitting the given architecture on the entire training set. Once re-trained, these were taken as the final classifiers for cross-domain testing.

4 Discussion

The resulting authorship attribution and gender profiling models were then deployed for validation on the Facebook caption test sets. The F1 score was employed since it accounts for false positives

Mixing(T:I)	Model	Attribution	Profiling
1:1	LR	0.650 ± $[0.517, 0.783]$	0.727 ± $[0.586, 0.840]$
2:0	LR	0.600 ± $[0.483, 0.717]$	0.714 ± $[0.560, 0.836]$
1:1	SVM	0.750 ± $[0.633, 0.867]$	0.697 ± $[0.542, 0.806]$
2:0	SVM	0.617 ± $[0.500, 0.733]$	0.655 ± $[0.500, 0.780]$
M.C.*	SVM & LR	0.333	0.500

Table 1: Model Evaluation. F1 scores for the final SVMs and LRs on the Facebook caption test sets. The models trained with a 1:1 mix of Twitter (T) and Instagram (I) captions performed 6% better on average in comparison to the pure Twitter models. *M.C. stands for the Majority Class baseline, as measured by accuracy.

and false negatives in a balanced manner, with the micro-averaged F1 taken for our multi-class authorship attribution task (*n*=3 classes). Confidence intervals on the F1 scores were then estimated with bootstrap resampling by taking 1000 samples with replacement from each model's test predictions and recalculating the F1 score against the ground truth labels per sample. The modeling metrics are shown above in Table 1. These results reveal significant improvement in F1 performance for the two cross-domain authorship analysis tasks explored, with mixed-platform models performing roughly 6% better on average compared with otherwise identical classifiers trained exclusively from Tweets. The SVMs benefited the most when trained with more platform-diverse datasets (13.3% F1 score improvement in authorship attribution, and 4.2% F1 score improvement in gender profiling), while LRs benefited less in comparison (5% F1 score improvement in authorship attribution, and 1.3% F1 score improvement in gender profiling).

The statistical significance for the experiments was then estimated with one-sided Welch's T-tests, which were taken over the medians of the resampled F1 scores. Our null hypothesis was that the difference in F1 performance between the single-platform and platform-mixed models was simply due to chance. Our alternative hypothesis was that the improvement in performance from multi-platform mixing was due to these models being trained from more heterogeneous training sets, as sourcing captions from multiple social media platforms enables the models to learn from different types of natural language. We obtained statistical significance in all four experiments, as adjusted via the Bonferroni correction to account for the four experimental tests performed. Three of the four tests yielded p-values below 1e-16, while the gender profiling LR presented a p-value of roughly 3e-3. And, in terms of practical significance, the experiments yielded Cohen's d effect sizes ranging from small to large (2.3, 0.83, 0.52, 0.24), with all but the gender profiling LR test exhibiting medium to large effect sizes.

5 Conclusions

The study's results help set concrete expectations for the benefits of using more platform-diverse datasets in NLP tasks on social media corpora. We present these findings to encourage future research on multi-platform mixing and domain adaptation in general, particularly when the modeling task at hand is unprecedented or intended for deployment under domain shift or on production data. While our results yield a promising snapshot of the value of multi-platform mixing, future research is needed to confirm these findings for other modeling tasks, different sized datasets, and to different degrees of domain shift.

6 Acknowledgements

We are grateful to Prof. David Bamman for discussions on domain adaptation in NLP and the mentorship of Prof. Duygu Ataman, which greatly improved the clarity and content of the research.

7 Software and Data

All code and data required to replicate the study are included in the corresponding zipped files. Modeling cards and data sheets further detailing our multi-platform mixing experiments are also contained therein at the "documentation cards" folder.

References

Antonio Theophilo Anderson Rocha, Efstathios Stamatatos. 2016. Authorship attribution for social media forensics. volume 12, pages 3–4.

Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. 2022. data2vec: A general framework for self-supervised learning in speech, vision and language.

Sabur Butt, Noman Ashraf, Grigori Sidorov, and Alexander Gelbukh. 2021. Sexism identification using bert and data augmentation – exist2021.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don't stop pretraining: adapt language models to domains and tasks. *arXiv preprint arXiv:2004.10964*.

Xiaochuang Han and Jacob Eisenstein. 2019. Unsupervised domain adaptation of contextualized embeddings for sequence labeling. *arXiv preprint arXiv:1904.02817*.

Mike Kestemont, Efstathios Stamatatos, Enrique Manjavacas, Walter Daelemans, Martin Potthast, and Benno Stein. 2019. Overview of the cross-domain authorship attribution task at pan 2019. In *CLEF (Working Notes)*.

Yawei Luo, Liang Zheng, Tao Guan, Junqing Yu, and Yi Yang. 2019. Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation.

D. M. Witten M. L. Jockers. 2010. A comparative study of machine learning methods for authorship attribution. volume 25, pages 215–223.

Jonas Peters, Dominik Janzing, and Bernhard Schlkopf. 2017. *Elements of Causal Inference: Foundations and Learning Algorithms*. The MIT Press.

Barbara Plank. 2016. What to do about non-standard (or non-canonical) language in nlp.

Victor Radivchev, Alex Nikolov, and Alexandrina Lambova. 2019. Celebrity Profiling using TF-IDF, Logistic Regression, and SVM—Notebook for PAN at CLEF 2019. In *CLEF 2019 Labs and Workshops, Notebook Papers*. CEUR-WS.org.

Alan Ramponi and Barbara Plank. 2020. Neural unsupervised domain adaptation in NLP—A survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6838–6855, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Yunita Sari, Mark Stevenson, and Andreas Vlachos. 2018. Topic or style? exploring the most useful features for authorship attribution. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 343–353, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Marcin Skowron, Marko Tkalčič, Bruce Ferwerda, and Markus Schedl. 2016. Fusing social media cues: Personality prediction from twitter and instagram. In *Proceedings of the 25th International Conference Companion on World Wide Web*, WWW '16 Companion, page 107–108, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Matti Wiegmann, Benno Stein, and Martin Potthast. 2019. Celebrity profiling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2611–2618, Florence, Italy. Association for Computational Linguistics.

Hongyi Zhang, Moustapha Cissé, Yann Dauphin, and David Lopez-Paz. 2018. mixup: Beyond empirical risk minimization. *ArXiv*, abs/1710.09412.

A Appendix

A.1

Instance weighting is the practice of re-weighting the estimate for the probability of any individual prediction by the probability of a document appearing in the target domain divided by the probability of a document appearing in the source domain. Instance weighting therefore assumes that the probability of a label is the same for a given document in both the source and target domains, a potentially unrealistic assumption. Feature augmentation is used when a large, labeled source domain exists to predict inference on a smaller, labeled target domain. Feature augmentation is the method of augmenting data representations to include multiple copies of differing form to the training set, including creative approaches such as the MixUp augmentation (Zhang et al., 2018). However, it is common to have a larger, unlabelled target domain, a situation where feature augmentation cannot be used. In these cases, domain-adversarial neural networks can be used to learn a representation for a document to predict the class label accurately and the domain (target v. source) inaccurately, and thus boosting generalization. Lastly, the method of pretraining involves taking a previously trained model and fine-tuning the model on data from the target domain (AdaptaBERT, DAPT) (Han and Eisenstein, 2019)(Gururangan et al., 2020).

A.2

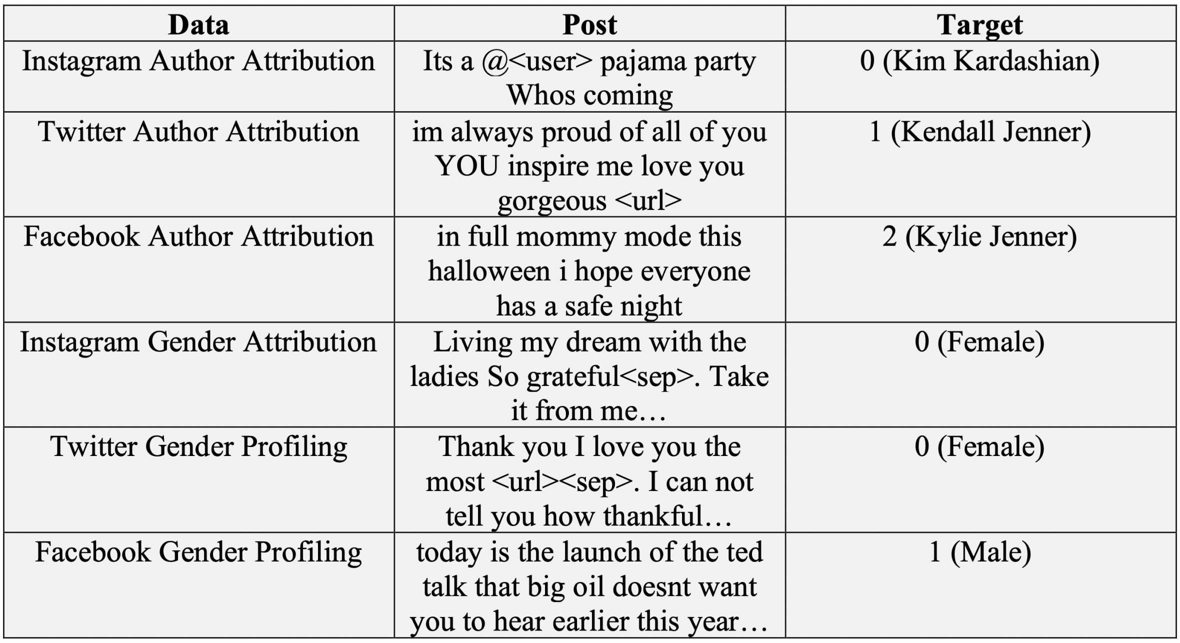
See Figure 2 below.

A.3

For authorship attribution, captions from Kylie Jenner, Kendall Jenner, and Kim Kardashian's Instagram, Twitter, and Facebook accounts were collected via scraping 62 captions per Instagram profile, 200 Tweets per Twitter profile, and 20 captions per Facebook profile. This data was then cleaned by replacing uninformative captions, which were defined as those with less than 2 distinct words, with random additional selections from the given account. In contrast, all 237 accounts were used for gender profiling, scraping up to 62 captions per Instagram profile (for a total of 11,594 captions), up to 200 Tweets per Twitter profile (for a total of 37,400 captions), and 20 captions per Facebook profile (totaling 1000 captions). The data was then pre-processed by removing all retweets and ads, tokenizing (with the NLTK punkt tokenizer), removing all symbols except for letters, numbers, at symbols, and hashtags, and replacing hyperlinks and user tags with 'url' and 'user' tokens respectively. This minimalist pipeline was inspired by the winner's of PAN18's authorship profiling challenge (Radivchev et al., 2019). Captions were piped individually for authorship attribution while they were concatenated for gender profiling per user with “sep“ tokens. The captions were then separated into training and testing sets. Each celebrity's Twitter and Instagram account contributed a row respectively to the mixed-platform training sets while each celebrity's Twitter account contributed two rows to the single-platform training sets. The number of captions concatenated per celebrity was held equivalent. In most cases 62 Instagram captions were scraped, and therefore, 62 x 2 Tweets were extracted from the scraped Twitter data for subsequent concatenation to two strings. In the under 5 percent of cases where less than 62 Instagram captions were scraped, the corresponding number of available Instagram captions y was used for determining the number of Twitter captions to concatenate per celebrity, yielding y x 2 Tweets in such cases.

A.4

BERT embedding features were also extracted for authorship attribution, which improved modeling test-set performance by roughly 7% relative to models trained solely with TF-IDF features. These were synthesized by taking BERT embeddings from the final layer of the Base-Cased network and averaging across token embeddings per post. These 768-dimensional contextual representations of the natural language captions were then merged to the TF-IDF features for Authorship Attribution modeling.



Data	Post	Target
Instagram Author Attribution	Its a @<user> pajama party Whos coming	0 (Kim Kardashian)
Twitter Author Attribution	im always proud of all of you YOU inspire me love you gorgeous <url>	1 (Kendall Jenner)
Facebook Author Attribution	in full mommy mode this halloween i hope everyone has a safe night	2 (Kylie Jenner)
Instagram Gender Attribution	Living my dream with the ladies So grateful<sep>. Take it from me…	0 (Female)
Twitter Gender Profiling	Thank you I love you the most <url><sep>. I can not tell you how thankful…	0 (Female)
Facebook Gender Profiling	today is the launch of the ted talk that big oil doesnt want you to hear earlier this year…	1 (Male)

Figure 2: Samples of the processed captions for each of the six dataset types. The authorship attribution posts are taken at the individual level, while the gender profiling posts are concatenated with “sep“ tokens (these are cut off for brevity in the figure above).